Method and Apparatus for Controlling Data Rat of a R v rs Link in a Communication System

Claim of Priority under 35 U.S.C. §119

[1001] The present Application for Patent claims priority to Provisional Application No. 60/448,269 entitled "Reverse Link Data Communication" filed February 18, 2003, and Provisional Application No. 60/469,376 entitled "Method and Apparatus for Controlling Data Rate of a Reverse Link in a Communication System" filed May 9, 2003, and assigned to the assignee hereof and hereby expressly incorporated by reference herein.

Field

[1002] The present invention relates generally to the field of communications, and more particularly, to controlling data rate of a reverse link from a mobile station in a communication system.

Background

[1003] In a wireless communication system, unnecessary and excessive transmissions by a user may cause interference for other users in addition to reducing the system capacity. The unnecessary and excessive transmission may be caused by inefficient selection of data rate of a reverse link in the communication system. The data communicated between two end users may pass through several layers of protocols for assuring proper flow of data through the system. Normally, a mobile station receives blocks of data from an application for transmission on a reverse link. The block of data is divided into a number of frames and transmitted over the communication link. The proper

delivery of data in at least one aspect is assured through a system of checking for error in each frame of data, and requesting a retransmission of the same frame of data if an unacceptable error or error rate is detected in the frame of data. The blocks of data may be of any type, for example, music data, video data, etc. The blocks of data may have different size and different delivery requirements. Such data delivery requirements often are associated with a quality of service. The quality of service may be measured by the communication data rate, the rate of packet loss that may be acceptable to the service, consistency in time delay of the data delivery, and an acceptable maximum delay for the communication of the data. Very often, if the data rate selected for transmission is not adequate, the required packet loss and the communication delay parameters may not be achieved.

[1004] On a forward link communication, the base station very often has adequate information about forward link quality with a number of mobile stations. As such, the base station may be able to centrally manage the forward link communication data rates. However, on the reverse link, a mobile station has no information about the transmissions from other mobile stations. Therefore, the mobile station may make a request to get permission to transmit at a data rate. The base station after reviewing every mobile station requests, accepts or rejects the requested data rate. If the requested data rate is rejected, the mobile station may request a lower data rate until the base station accepts a requested data rate. The mobile station may have permission to transmit below a data rate without going through the request and acceptance process. Such a data rate is normally a very low data rate. Before transmission on the reverse link, the mobile station needs to have completed its

communication for the data rate request. Such overhead communications between the mobile stations and the base stations may reach an unacceptable level and impact the desired quality of service.

[1005] Therefore, there is a need to provide a system, method and apparatus for selection of a reverse link data rate in a communication system.

BRIEF DESCRIPTION OF THE DRAWINGS

[1006] The features, objects, and advantages of the present invention will become more apparent from the detailed description set forth below when taken in conjunction with the drawings in which like reference characters identify correspondingly throughout and wherein:

[1007] FIG. 1 depicts a communication system for transmitting and receiving data in accordance with various aspects of the invention;

[1008] FIG. 2 depicts a receiver system for receiving data in accordance with various aspects of the invention;

[1009] FIG. 3 depicts a transmitter system for transmitting data in accordance with various aspects of the invention; and

[1010] FIG. 4 depicts a flow of messages and processes for determining a data rate for reverse link communication.

Detailed Description of the Preferred Embodiment(s)

[1011] One or more exemplary embodiments described herein are set forth in the context of a digital wireless data communication system. While use within this context is advantageous, different embodiments of the invention may be incorporated in different environments or configurations. In general, the

various systems described herein may be formed using software-controlled processors, integrated circuits, or discrete logic. The data, instructions, commands, information, signals, symbols, and chips that may be referenced throughout the application are advantageously represented by voltages, currents, electromagnetic waves, magnetic fields or particles, optical fields or particles, or a combination thereof. In addition, the blocks shown in each block diagram may represent hardware or method steps.

[1012] More specifically, various embodiments of the invention may be incorporated in a wireless communication system operating in accordance with the code division multiple access (CDMA) technique which has been disclosed and described in various standards published by the Telecommunication Industry Association (TIA) and other standards organizations. Such standards include the TIA/EIA-95 standard, TIA/EIA-IS-2000 standard, IMT-2000 standard, UMTS and WCDMA standard, all incorporated by reference herein. A system for communication of data is also detailed in the "TIA/EIA/IS-856 cdma2000 High Rate Packet Data Air Interface Specification," incorporated by reference herein. A copy of the standards may be obtained by accessing the world wide web at the address: http://www.3gpp2.org, or by writing to TIA, Standards and Technology Department, 2500 Wilson Boulevard, Arlington, VA 22201, United States of America. The standard generally identified as UMTS standard, incorporated by reference herein, may be obtained by contacting 3GPP Support Office, 650 Route des Lucioles-Sophia Antipolis, Valbonne-France.

[1013] FIG. 1 illustrates a general block diagram of a communication system 100 capable of operating in accordance with any of the code division

multiple access (CDMA) communication system standards while incorporating various embodiments of the invention. Communication system 100 may be for communications of voice, data or both. Generally, communication system 100 includes a base station 101 that provides communication links between a number of mobile stations, such as mobile stations 102-104, and between the mobile stations 102-104 and a public switch telephone and data network 105. The mobile stations in FIG. 1 may be referred to as data access terminals (AT) and the base station as a data access network (AN) without departing from the main scope and various advantages of the invention. Base station 101 may include a number of components, such as a base station controller and a base transceiver system. For simplicity, such components are not shown. Base station 101 may be in communication with other base stations, for example base station 160. A mobile switching center (not shown) may control various operating aspects of the communication system 100 and in relation to a back-haul 199 between network 105 and base stations 101 and 160.

[1014] Base station 101 communicates with each mobile station that is in its coverage area via a forward link signal transmitted from base station 101. The forward link signals targeted for mobile stations 102-104 may be summed to form a forward link signal 106. The forward link may carry a number of different forward link channels. Each of the mobile stations 102-104 receiving forward link signal 106 decodes the forward link signal 106 to extract the information that is targeted for its user. Base station 160 may also communicate with the mobile stations that are in its coverage area via a forward link signal transmitted from base station 160. Mobile stations 102-104 may communicate with base stations 101 and 160 via corresponding reverse links.

Each reverse link is maintained by a reverse link signal, such as reverse link signals 107-109 for respectively mobile stations 102-104. The reverse link signals 107-109, although may be targeted for one base station, may be received at other base stations.

[1015] Base stations 101 and 160 may be simultaneously communicating to a common mobile station. For example, mobile station 102 may be in close proximity of base stations 101 and 160, which can maintain communications with both base stations 101 and 160. On the forward link, base station 101 transmits on forward link signal 106, and base station 160 on the forward link signal 161. On the reverse link, mobile station 102 transmits on reverse link signal 107 to be received by both base stations 101 and 160. For transmitting a packet of data to mobile station 102, one of the base stations 101 and 160 may be selected to transmit the packet of data to mobile station 102. On the reverse link, both base stations 101 and 160 may attempt to decode the traffic data transmission from the mobile station 102. The data rate and power level of the reverse and forward links may be maintained in accordance with the channel condition between the base station and the mobile station as outlined by various aspects of the invention.

[1016] FIG. 2 illustrates a block diagram of a receiver 200 used for processing and demodulating the received CDMA signal while operating in accordance with various aspects of the invention. Receiver 200 may be used for decoding the information on the reverse and forward links signals. Receiver 200 may be used for decoding information on the fundamental channel, control channel and supplemental channels. Received (Rx) samples may be stored in RAM 204. Receive samples are generated by a radio frequency/intermediate

frequency (RF/IF) system 290 and an antenna system 292. The RF/IF system 290 and antenna system 292 may include one or more components for receiving multiple signals and RF/IF processing of the received signals for taking advantage of the receive diversity gain. Multiple received signals propagated through different propagation paths may be from a common source. Antenna system 292 receives the RF signals, and passes the RF signals to RF/IF system 290. RF/IF system 290 may be any conventional RF/IF receiver. The received RF signals are filtered, down-converted and digitized to form RX samples at base band frequencies. The samples are supplied to a multiplexer (mux) 252. The output of mux 252 is supplied to a searcher unit 206 and finger elements 208. A control system 210 is coupled thereto. A combiner 212 couples a decoder 214 to finger elements 208. Control system 210 may be a microprocessor controlled by software, and may be located on the same integrated circuit or on a separate integrated circuit. The decoding function in decoder 214 may be in accordance with a turbo decoder or any other suitable decoding algorithms. The signal transmitted from a source may be encoded with several layers of codes. The decoder 214 may perform decoding function in accordance with two or more codes. For example, the transmitted data may be encoded at two different layers, an outer layer and a physical layer. The physical layer may be in accordance with the Turbo code, and the outer layer may be in accordance with Reed Solomon code. As such, the decoder 214 decodes the received samples in accordance with such codes.

[1017] During operation, received samples are supplied to mux 252. Mux 252 supplies the samples to searcher unit 206 and finger elements 208. Control unit 210 configures finger elements 208 to perform demodulation and

despreading of the received signal at different time offsets based on search results from searcher unit 206. The results of the demodulation are combined and passed to decoder 214. Decoder 214 decodes the data and outputs the decoded data. Despreading of the channels is performed by multiplying the received samples with the complex conjugate of the PN sequence and assigned Walsh function at a single timing hypothesis and digitally filtering the resulting samples, often with an integrate and dump accumulator circuit (not shown). Such a technique is commonly known in the art. Receiver 200 may be used in a receiver portion of base stations 101 and 160 for processing the received reverse link signals from the mobile stations, and in a receiver portion of any of the mobile stations for processing the received forward link signals.

[1018] The decoder 214 may accumulate the combined energy for detection of a data symbol. Each packet of data may carry a cyclic redundancy check (CRC) field. The decoder 214 may in connection with control system 210 and or other control systems check for error in the received data packet. If the CRC data does not pass, the received packet of data has been received in error. The control system 210 and or other control systems may send a negative acknowledgment message to the transmitter to retransmit the packet of data.

[1019] FIG. 3 illustrates a block diagram of a transmitter 300 for transmitting the reverse and forward link signals. The channel data for transmission are input to a modulator 301 for modulation. The modulation may be according to any of the commonly known modulation techniques such as QAM, PSK or BPSK. Before modulation, the channel data for transmission may pass through one or more layers of coding. The channel data for transmission

are produced for modulator 301. The channel data for transmission are received by the modulator 301.

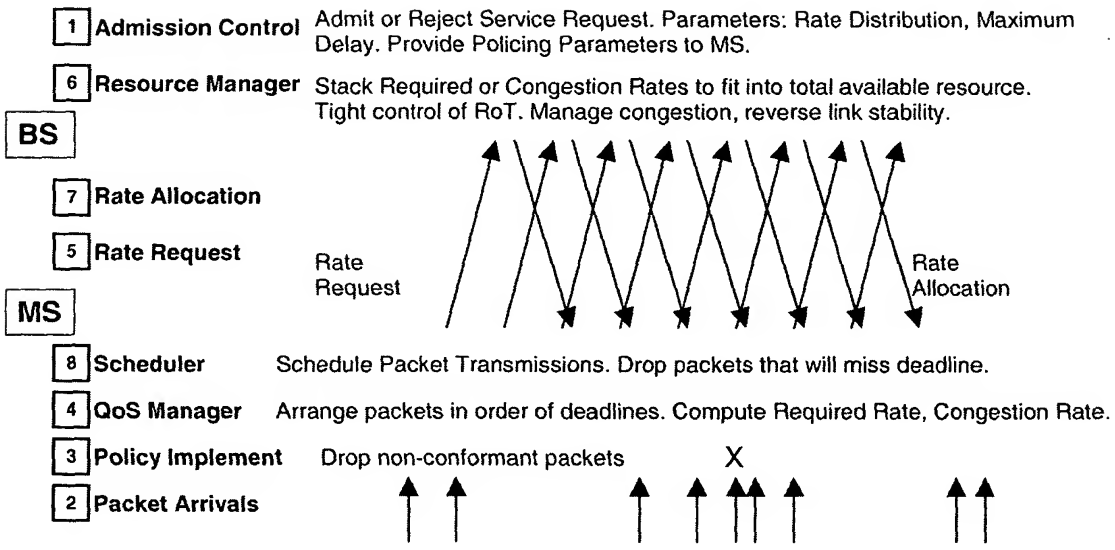
[1020] The modulation data rate may be selected by a data rate and power level selector 303. The data rate selection may be based on feedback information received from a destination. The data rate very often is based on the channel condition, among other considered factors and in accordance with various aspects of the invention. The channel condition may change from time to time. The data rate selection may also change from time to time.

[1021] The data rate and power level selector 303 accordingly selects the data rate in modulator 301. The output of modulator 301 passes through a signal spreading operation and amplified in a block 302 for transmission from an antenna 304. The data rate and power level selector 303 also selects a power level for the amplification level of the transmitted signal. The combination of the selected data rate and the power level allows proper decoding of the transmitted data at the receiving destination. A pilot signal is also generated in a block 307. The pilot signal is amplified to an appropriate level in block 307. The pilot signal power level may be in accordance with the channel condition at the receiving destination. The pilot signal may be combined with the channel signal in a combiner 308. The combined signal may be amplified in an amplifier 309 and transmitted from antenna 304. The antenna 304 may be in any number of combinations including antenna arrays and multiple input multiple output configurations.

[1022] In CDMA 2000 system, the mobile station (MS) is permitted to have several simultaneous communication services. Each one of the communication services may have different quality of service (QoS)

requirements. For a service option, packets of data may be communicated with the specifically defined QoS parameters such as a specific data rate or a range of data rates, a packet loss rate and a maximum delay allowed for communication of data packet or a number of data packets. During the service negotiation phase of a communication link, MS and the base station (BS) agree on a set of QoS parameters. The QoS parameters may be defined for duration of a defined communication service. The BS then may be required to meet such negotiated QoS such as the date rate, packet loss and maximum delay with high probability.

In accordance with various aspects of the invention, a method and apparatus provide for an implementation of a QoS on the reverse link, where an updated information relating to the queue length and packet delay deadlines are available at the MS, while the resource manager allocating the negotiated QoS is at the BS. The MS requests a required rate from the BS rather than reporting its queue length (backlog) information. The MS computes the required data rate and duration before requesting the data rate from the BS. The request for the data rate may be in the form of requesting one or more forward link traffic channel power to pilot (T/P) ratio. The set of available data rates may have corresponding T/P ratios. A table may provide for the correspondence between the T/P and the data rate. The autonomous data rate control by the MS may also be based on a congestion feedback from the BS. The BS may be responsible for allocating rate to the MS's and for congestion management and stability of the reverse link. The BS is also responsible for admission control. The allocation of resources in response to a data rate request may be depicted in the following graphical flow of messages and processes.



[1023] The actual resource managed by the BS is the traffic channel to pilot power ratio (T/P). The mapping from data rate to T/P of the channel is the operating point that is chosen based on the number of permitted retransmissions and the use of hybrid ARQ. The BS may assign a different mapping as a function of the delay requirement (permitted retransmissions) for each service. Such an optimization is useful for services with short transactions and very low delay requirements (e.g., interactive gaming). For most services, the best choice is for the BS to choose the mapping that maximizes reverse link throughput. The numbers (1-8) on the left hand side represent a possible order of events or processes that may take place.

1. The BS manages admission control and only admits the communication service (or flows) with acceptable and achievable QoS requirements. Once a service or communication flow of data packets is admitted, the MS is aware of the negotiated QoS parameters, such as acceptable data

rate, packet loss rate and maximum delay associated with the flow. Note that due to channel variations and mobility, these QoS guarantees are necessarily probabilistic.

2. The MS implements an (upstream) policy that discards non-conformant data packets at the ingress. Thus, the MS accepts all packets that by the implementation of the policy are assumed to meet the negotiated QoS for the admitted flows. The packets that require a QoS exceeding the negotiated QoS as defined by for example communication data rate are discarded at the MS prior to the output queuing stage. The MS may also implement an outer-loop mechanism to adjust the policing parameters based on operating conditions. The BS may "verify" that the MS is actually conforming to its negotiated rate.
3. Conformant packets admitted at the MS are placed in the output queue. A deadline is associated with each packet based on the packet arrival time and the maximum permitted delay for that service (or flow). Preferably, the MS may arrange the output queue so that the packets are stored in the order of their deadlines, the earliest deadline first. The MS must manage its transmission schedule to ensure that packets are transmitted before their deadline.
4. The MS determines a required data rate based on the deadlines associated with the packets in the output queue. The procedure is described below more fully. Since the data rate determined by the MS is required to meet the negotiated QoS, the requested data rate is not just a "priority" indication. The MS also computes one or more congestion data rates by determining what packets in its queue may be dropped to assist

the BS in making an allocation lower than the required rate, if the required rate cannot be allocated due to congestion, overload control or any other reason, by the base station. The congestion rates are determined by the MS based on the required rate and the number of packets that may be dropped in the queue. Generally, when a packet is dropped from the queue, the data rate needed to transmit the remaining number of packets decreases. The BS translates the required and congestion rates to the required and congestion T/P, or alternately the MS can directly compute a required and congestion T/P.

5. The MS communicates the required and congestion T/P or T/P increment or decrement and communicates it to the BS. The BS must attempt to meet requests for T/P increases subject to available resources since these resources are required by the MS to meet its QoS criteria. Several consecutive T/P increase requests from the MS indicates increasing priority, which if not satisfied would result in some QoS criteria not being satisfied.
6. The BS scheduler stacks these T/P requests in terms of reverse link resources (i.e., rise over thermal) and time. The BS also sets aside certain resources for known low delay, constant bandwidth flows e.g., voice calls. The BS may attempt to optimize such a stacking, e.g., by delaying certain T/P allocations or providing higher T/P allocation for shorter duration. If the BS delays an allocation to the MS, subsequent requests from the MS may request even higher T/P to satisfy the QoS for delay-sensitive packets, because the longer the delay in transmitting a data packet, the higher the date rate required for transmission to meet

the same QoS. Thus, the BS has limited flexibility in scheduling. If there is an excess bandwidth available, the BS may choose to ignore requests for T/P decreases, or provide higher than requested T/P.

7. The BS makes T/P allocation to the MS. The allocation may be indicated to the MS as an increment (or decrement) to a current data rate allocation.
8. Based on the T/P allocation, the MS schedules packets for transmission. The MS serves packets based on earliest-deadline-first scheduling discipline, and may be subject to a modification. For example, prior to starting the transmission of any packet, the MS determines if the transmission of the data packet takes place within its deadline. This determination is a function of the allocated T/P and the deadline and should account for the potential for an increased allocation in the future. The MS drops any packet that is likely to miss its transmission deadline. A packet that is not successfully transmitted prior to the expiry of its deadline is counted as lost. The MS tracks the packet loss rate associated with the flow.

[1024] In such a framework, the processes of time steps 2, 3 and 8, respectively, allow the MS to manage the QoS (rate, maximum delay and packet loss guarantees) associated with its flows. The processes of time steps 4 and 5 allow the MS to merge its needs for all its flows into one T/P requirement. The BS admission control process at time step 1 ensures that the BS would have enough resources in time step 6 to satisfy the requirements of all admitted, negotiated QoS flows from all MS's. The MS determines the required data rate to satisfy the QoS. The MS merges queues for multiple

(negotiated QoS) services into a rate requirement. Moreover, generally, instead of determining required rate and translating to a required T/P, the MS can work directly with required T/P. This is more general, since it easily accommodates the transmission of packets from different services with different mappings from T/P to rate.

[1025] Let us assume that at time t_0, the MS queues consist of packets P_i, {i=1, ..., N} of size s_i, arranged in order of their deadlines d_i. Each packet P_i is associated with a data service k(i). For data service k, the known data rate to T/P map is defined as: $R_k(T/P)$. Then, at the allocated T/P value T_0, the following equations may be defined and determined. At rate $R_{k(i)}(T_0)$, the transmission time of packet P_i over the air, x_i is

$$x_i = s_i / R_{k(i)}(T_0), \quad (1)$$

Since the packets have been arranged in order of their deadlines, packet P_i will complete its transmission at

$$z_i = t_0 + \sum s_j / R_{k(j)}(T_0), \text{ where the sum is over } [1, ..., i]. \quad (2)$$

That is, packets P_1, ..., P_{i-1}, with deadlines prior to d_i are transmitted before P_i. Thus, the process may determine if any packet in the MS output queue would miss its deadline, i.e.

$$z_i > d_i, \text{ for } 1 \leq i \leq N \quad (3)$$

If the MS determines that any packet in its queue would miss its deadline at a current rate, then it may require a higher rate to satisfy its QoS. Note: such a data rate computation uses the deadline information associated with every packet in the MS queue. The BS is unable to make such a data rate determination based on just the backlog and QoS class.

[1026] There may be several ways of providing the required T/P information to the BS. Depending on the design of the request channel and the frequency of transmission of request information from the MS to BS, it may also be useful to compute a required duration at the MS and provide an indication to the BS. The above equation 2 also allows the MS to determine the required duration. Basically, at the allocated T/P ratio T_0, the last packet in the MS queue would complete its transmission at time z_N. Therefore, based on the current packet queue and the allocated rate, the required duration is z_N. The equation 2 may be updated with very little computation burden. For example, if at a subsequent time t_1 at the completion of the transmission of packet P_1, the allocated T/P has changed to T_1, the updated packet completion instants are written as:

$$z_i(t_1) = t_1 + \sum s_j / R_{k(j)}(T_1), \text{ where the sum is over } [2, \ldots, i]. \quad (4)$$

These completion times can be computed from the previous packet completion times using the equation:

$$z_i(t_1) - t_1 = z_i(t_0) - t_0 - s_1/R_{k(1)}(T_0) + \sum s_j \left[1/R_{k(j)}(T_1) - 1/R_{k(j)}(T_0) \right] \quad (5)$$

Next, assume that a new packet P_{new} of size s_{new}, service k(new) and deadline d_{new} arrives at time t_2. In general, the deadline for the new packet is between the (ordered) deadlines for packets k and k+1, i.e. for some $k < N$, $d_k \leq d_{new} < d_{k+1}$. Then, for $i \leq k$, $z_i(t_2)$ is unchanged. If the T/P value T_1 is unchanged,

$$z_i(t_2) = z_i(t_1) + s_{new}/R_{k(new)}(T_1) \quad \text{for } i > k \quad (6)$$

Thus, the MS is able to compute and keep updated its required T/P, required duration as well as the transmission schedule of packets in its queue.

[1027] To assist the BS in making rate allocations during periods of congestion, the MS also computes congestion rates by determining what

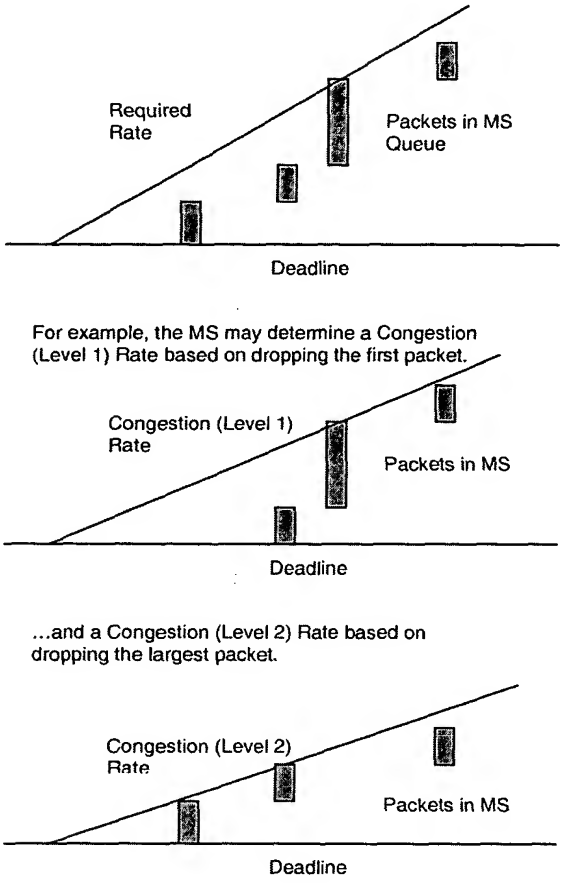
packets in its queue may be dropped. The MS may use many criteria to prioritize which packets can be dropped:

- Packets from services that are tolerant to dropped packets,
- Packets from services whose current packet loss rate is smaller than the negotiated packet loss rate,
- Packets that are likely to miss their delay deadline if the required T/P is not allocated.

[1028] Based on the drop priority, the MS determines which packets can potentially be dropped while still satisfying an acceptable QoS at various levels of congestion. The MS then applies equations (1) through (3) to a virtual queue formed by removing these packets from the MS queue to compute the congestion T/P values. Note, throughout, the packet data and block of data may be interchangeable.

[1029] If the T/P to Rate mapping is fixed, it is more convenient and equivalent to work with data rate. A schematic that illustrates graphically, the required and congestion data rate computations is shown below. The packet sizes and the deadlines are also graphically shown. The size of each packet (in bits) is shown as a vertical bar placed at its deadline in a time frame axis. The size of the vertical bar is a representative of the packet size. Any line with positive slope pivoted at the origin corresponds to a rate (in bits/second). The origin is the current time or time of start of allocation. The required rate is the smallest slope that satisfies all deadlines, i.e., the smallest slope such that all the packets in the chart are below the line. Also, where the MS computes the congestion rate assuming that the first packet in the queue can be dropped

(congestion level 1) and a congestion rate associated with dropping the largest packet in the queue (congestion level 2).



At time t_0, the MS queues consist of packets P_i, {i=1, ..., N} of size s_i, arranged in order of their deadlines d_i. Then, at the allocated rate R_0 we can write the following equations. At rate R_0, the transmission time of packet P_i over the air, x_i is

$$x_i = s_i / R_0 \qquad (7)$$

Since the packets have been arranged in order of their deadlines, packet P_i will complete its transmission at

$$z_i = t_0 + \sum s_j / R_0 \text{ where the sum is over } [1, \ldots, i]. \qquad (8)$$

That is, packets P_1, ..., P_{i-1}, with deadlines prior to d_i are transmitted before P_i. Thus, the process may determine if any packet in the MS output queue would miss its deadline, i.e.

$$z_i > d_i \text{ , for } 1 \leq i \leq N \qquad (9)$$

If the MS determines that any packet in its queue would miss its deadline, then it requires a higher rate to satisfy its QoS. Such a data rate computation uses the deadline information associated with every packet in the MS queue. The BS is unable to make this rate determination based on just the backlog and QoS class.

There are several ways of providing the required rate information to the BS. Depending on the design of the request channel and the frequency of transmission of request information from the MS to BS, it may also be useful to compute a required duration at the MS and provide an indication to the BS. The above (equation 2) also allows the MS to determine the required duration. Basically, at the allocated rate R_0, the last packet in the MS queue would complete its transmission at time z_N. Therefore, based on the current packet queue and the allocated rate, the required duration is z_N. The equation (2) may be updated with very little computation burden. For example, if at a subsequent time t_1 at the completion of the transmission of packet P_1, the allocated rate has changed to R_1, the updated packet completion instants are written as:

$$z_i(t_1) = t_1 + \sum s_j/R_1 \text{ where the sum is over } [2, \ldots, i]. \qquad (10)$$

These completion times can be computed from the previous packet completion times using the equation:

$$z_i(t_1) - t_1 = [(z_i(t_0) - t_0) R_0 - s_1]/R_1 \qquad (11)$$

Next, assume that a new packet P_{new} of size s_{new} and deadline d_{new} arrives at time t_2. In general, the deadline for the new packet may be between the (ordered) deadlines for packets k and k+1, i.e. for some $k < N$, $d_k \leq d_{new} < d_{k+1}$. Then, for $i \leq k$, $z_i(t_2)$ is unchanged. If the rate R_1 is unchanged,

$$z_i(t_2) = z_i(t_1) + s_{new}/R_1 \quad \text{for } i > k \quad (12)$$

Thus, the MS is able to compute and keep update its required rate, required duration as well as the transmission schedule of packets in its queue.

[1030] As shown above, by examining the delay deadlines of all the packets in the MS buffer, the MS is able to determine a required rate or T/P. Alternatively, if the MS examines only the first packet in its buffer, that is, the packet with the shortest delay deadline, and applies the T/P (or rate) calculations described above, the result is equivalent to delay deadline feedback to the BS. In this case, the value of calculated T/P (or rate) represents the shortest deadline and equivalently highest priority. An example of a two-bit encoding of delay deadlines for the reverse link channel is shown later paragraphs and may be used.

[1031] The reverse link requests may be sent to the BS either through messages or using a continuous low-rate rate control channel. The following schemes may be considered and used:

- Using a reverse link message to provide queue length or backlog information to the BS. For QoS support, a QoS field may be added to this request message.
- A continuous T/P or rate request channel where the MS periodically inserts a bit indicating a request for higher rate. This does not provide the BS any indication of QoS.

The resources managed by the BS include the traffic channel to pilot power ratio (T/P). Generally, a higher T/P ratio maps to a higher data rate. The system may allow more than one mapping scheme between the T/P ratio and the corresponding data rate. In general, the MS always chooses the data rate to T/P mapping that maximizes reverse link throughput. For some services (e.g., interactive gaming) with short transactions and very low delay requirements, it may be necessary to operate with fewer retransmissions and a higher T/P. Thus, if the data packet at the head of the queue at the MS has a very short deadline (e.g., less than 40 ms), the MS may choose a special rate to T/P mapping suitable for low-delay services. It is possible to operate the scheme either using rate, and mapping it to T/P, or directly computing required T/P.

[1032] To allow the BS resource manager to prioritize allocation in periods of congestion, in addition to the required T/P, the MS also indicates one or more values of congestion T/P to the BS. The BS attempts to fairly allocate the required T/P to all MS. If the required resources exceeds the available resources, the BS changes the required T/P for one MS at a time (in some order determined by the BS), with a smaller T/P associated with congestion level 1 until the total allocation for all mobile stations falls within the available resources. If necessary, the BS may move on to the T/P associated with congestion level 2, etc. The indication of the required T/P or data rate and congestion T/P or rates may be communicated via short control messages, continuous messaging or a combination.

[1033] In the request message to the BS, the required T/P and its duration may be provided based on the computations at the MS. Such a

request message may not include the backlog and QoS feedback to the BS. The BS is not able to compute the required T/P and its duration based on the backlog and QoS class. To manage QoS (rate, packet loss, maximum delay), periodic messages that request T/P and its duration are preferable to periodic feedback of backlog and QoS class. In response, the BS makes a T/P and duration allocation to the MS through a grant message. The MS continues to update its (local) rate and duration computation. An updated request is triggered whenever there is either a significant change in the required T/P, or the required duration exceeds the allocated duration by a significant amount. A grant with a zero T/P allocation may indicate the termination of a grant to the MS.

[1034] Once a T/P is granted to the MS, to reduce request messaging overhead, a low-bandwidth continuous reverse link request channel may be utilized. The MS maintains a Current Grant variable based on the grant from the BS. Alternately, it is possible that the grant is implicit, that is, any MS is allowed to autonomously set its Current Grant variable to a global (initial) value of Current Grant and thus eliminate the need for any messages.

[1035] Based on the required rate computations, the MS continuously sends requests to increase, decrease or leave unchanged its Current Grant. A request for increased T/P may also indicate if the increase is needed to satisfy the T/P at various levels of congestion. An encoding of a 2-bit Differential Rate Request field may be included in the message. The MS indicates the level of its Current Grant with respect to the required rate and congestion rates, for example, if the Current Grant is between the required T/P and the congestion level 1 T/P, then the MS requests includes encoded bit "10". In an alternative, a

2-bit Differential Rate Request field that contains only 1 congestion level and a new level that prevents buffer underflow for delay sensitive traffic may be included.

	Current Grant	Request
		11
Required T/P		
		10
Congestion Level 1 T/P		
		01
Congestion Level 2 T/P		
		00

The T/P Request indicates the level of the Current Grant to the MS with respect to the Required T/P and the Congestion T/P.

Encoding of 2-bit Differential T/P Request Field

	Current Grant	Request
		11
T/P can be reduced		
		10
Required T/P		
		01
Congestion Level T/P		
		00

The T/P Request indicates the level of the Current Grant to the MS with respect to the Required T/P, the Congestion T/P and T/P that can create buffer underflow.

Alternative Encoding of 2-bit Differential T/P Request Field

[1036] Optionally, a look-ahead long grant threshold D_0 may also be defined. If the MS computes its required duration exceeds D_0, then it may indicate a request for a long grant. This is useful, as it allows the BS to look-ahead and thus better manage its resource allocation and scheduling decisions. For the scheduler, requests for high rate are "priority" requests, while requests for long grant indicate backlog.

[1037] Alternately, two bits may be used to represent the delay deadline (or priority level) of the packet at the head of the queue. For example:

Priority Level	Delay Deadline of head of queue packet
3	less than X
2	greater than X, but less than 3X
1	greater than 3X, but less than 9X
0	greater than 9X (i.e., best effort)

X is a system parameter whose value may be fixed for the BS, or may be fixed depending on the mix of services per MS.

[1038] The MS transmits on the reverse link using the T/P value Current Grant. The BS determines the value of the Current Grant variable at the MS from the current transmission. The BS may determine this value by measuring the ratio of the traffic channel power to the pilot power in the MS transmission, or from the rate used by the MS for the transmission and then mapping it to T/P.

[1039] The BS resource manager uses the current T/P used by the MS, along with the information in the T/P request to allocate T/P fairly among the MS. For example, depending on the level of congestion it may be able to satisfy only the congestion level 1 requirements for all MS. The BS would then allocate T/P increases for MS whose request indicates 00 or 01, with higher priority for MS requests indicating 00. The BS would allocate T/P decreases for MS whose request indicates 11 or 10. The BS may also use additional criteria to manage contention among the MS.

[1040] The BS resource manager operation may be explained with a following example. Considering the case of three MS:

MS 1: Ec_Pilot[1], Current Grant [1], Request = 10

MS 2: Ec_Pilot[2], Current Grant [2], Request = 01

MS 3: Ec_Pilot[3], Current Grant [3], Request = 01

MS 4: Ec_Pilot[4], Current Grant [4], Request = 11

MS 5: Ec_Pilot[5], Current Grant [5], Request = 00

MS 6: Ec_Pilot[6], Current Grant [6], Request = 10

Noting the above example, all MS except MS 4 require higher T/P than the Current Grant to meet the required T/P. The MS 4 may be assigned a decrease in its Current Grant. The allocation of increase may or may not be offered to other MS, depending on further computations. The BS resource manager is able to compute the current resource utilization from the Ec_Pilot and Current Grant for each MS, as follows:

$$\text{Resource Utilization} = \sum \text{Ec\_Pilot}[i]*[\text{Current Grant}[i] + 1] \quad (13)$$

[1041] The BS grants an increment (up adjust) or a decrement (down adjust) with respect to the Current Grant. The increment or decrement is a multiplicative factor to the Current Grant whose value Adjust[i] is given by a (1+a) or (1-a), respectively. Then following an allocation, the updated resource utilization can be computed as:

$$\text{Resource Utilization} = \sum \text{Ec\_Pilot}[i]*[\text{Adjust}[i]*\text{Current Grant}[i] + 1] \quad (14)$$

The resource manager algorithm may proceed through the following steps. It terminates at the step where a set of adjustment values Adjust[i] is found where the updated resource utilization is below the maximum resource utilization threshold Tmax.

Step 1 Assigning decrement to MS 4 and increment to all other MS. Adjust[4]=1-a, Adjust[i]=1+a, for i=1,2,3,5,6. This allocation attempts to move all MS to meet their required T/P. Determining the updated resource

utilization. If the updated resource utilization is below the maximum threshold Tmax, this allocation is permitted and may be assigned. If the resource utilization exceeds Tmax, then move to Step 2.

Step 2 It is not possible to move all MS towards the required T/P. For fairness, the process moves all MS towards the congestion level 1 T/P. This means that in addition to MS 4, MS 1 and 6 can be assigned a down adjust. That is, Adjust[i]=1-a, for i=1,4,6 and Adjust[i]=1+a for i=2,3,5. Once again, determine if this updated resource allocation is permitted or not, i.e. does not exceed the total allowed allocation. If not permitted, then move to Step 3.

Step 3 It is not possible to move all MS towards the congestion level 1 T/P. For fairness, the process moves all MS towards the congestion level 2 T/P. This means that all MS except MS 5 can be assigned a down adjust. That is, Adjust[i]=1-a, for i=1,2,3,4,6 and Adjust[5]=1+a. Once again, determining if this updated resource allocation is permitted or not i.e. does not exceed the total allowed allocation. If not permitted, then move to Step 4.

Step 4 The BS T/P adjustment algorithm is unable to determine a satisfactory allocation. An explicit message may be required to terminate the transmission from one or MS. The BS chooses which grant to terminate based on various criteria including: fairness and the size of the current allocation to the MS.

[1042] A low-bandwidth continuous forward link grant channel may be used to indicate adjustments to the Current Grant to the MS. The MS modifies its Current Grant variable based on the actual grants (and adjustments) it

receives from the BS. Since grants are continuously extended to the MS, it is not necessary to indicate a long grant to the MS. The long grant request (if used) just allows the BS to look-ahead and thus making better scheduling decisions.

[1043] The encoding on the low-bandwidth continuous forward link grant channel may be as follows:

+1 if the mobile station is instructed to increase T/P by configured amount, a (could be also T/P dependent)

-1 if the mobile station is instructed to decrease T/P by configured amount, a (could be also T/P dependent)

0 if the mobile station is instructed to keep T/P unchanged.

[1044] If the BS does not decode the continuous request channel reliably (for example the reverse pilot channel (R-PICH) is received with low power and continuous request channel symbols are erased), the BS sets the forward grant channel symbol to 0. Otherwise, if the continuous request channel is reliably decoded, the BS sets the forward grant channel symbol appropriately.

[1045] While the BS resource manager determines the congestion level based on the MS requests, and instead of requiring a continuous grant channel per MS, with some loss of flexibility, it is possible to use a continuous common grant channel that just indicates the current computed level of congestion at the BS (for example, encoded as one of three levels shown above). Based on such an indication, the MS can autonomously adjust its data rate following the same logic. For example, when the BS common grant indicates congestion level 1, MS requesting 10 must decrease their T/P while MS requesting 01 can increase their T/P for the subsequent transmission. Moreover, while reducing the

overhead on the forward link, the continuous common grant channel takes away the ability of the BS to discriminate among contending MS based on additional criteria. It is also possible to eliminate the continuous request channel and reduce this to autonomous operation by the MS. The congestion indication from the BS is based on current measurements of resource usage by the MS, rather than MS requests. Thus, there is no closed loop per MS. In this case, the BS is unable to distinguish between MS whose QoS is being satisfied and those whose QoS is not being met and so is unable to enforce fairness.

[1046] As such, signaling errors on the continuous request and grant channels are not catastrophic. An erasure on the continuous forward grant channel is assumed to be a command to leave the Current Grant unchanged. Due to the low-delay closed loop mechanism any signaling errors would quickly be corrected by subsequent requests and grants. In particular, the local variable Current Grant at the MS becomes known to the BS at every reverse link frame transmission and allows the two sides of the control loop to maintain the same state.

[1047] It is possible to operate the reverse link QoS management using just the request and grant messages. However, for tight QoS management, i.e. to manage QoS for services with bursty arrivals, variable rates and tight delay constraints, it is necessary to use the continuous request and grant channels. Note that the continuous request channel can be used without the continuous grant channel.

For a negotiated QoS traffic, the MS may operate as following:

- The MS sends a request message that indicates the required T/P (and congestion T/P). The request message may also contain maximum T/P

(the headroom at the MS). The BS may indicate a T/P grant through a message, or the grant may be implicit (a known initial T/P). In the latter case, no request or grant message is required.

- Changes to the required T/P are indicated on the low-overhead continuous rate request channel. The channel is subject to set up and release with Layer 3 signaling. Since this channel represents the aggregate rate requirement for the MS, only one such channel is needed per MS.
- Changes to the granted T/P are indicated by the base station on the low-overhead continuous grant channel on the forward link. Alternately, a low-overhead continuous common grant channel may be used which indicates the congestion level based on the received requests.

[1048] During a soft handoff operation, the message-based reverse resource management is handled by the serving BS only (the BS with the highest averaged pilot level received at the MS). The continuous request and grant channels may be received and transmitted only by the serving BS or by a reduced active set of BS. If the continuous grant channel is transmitted only by the serving BS, the operation in soft handoff does not differ from the case when mobile station is not in soft handoff. However, if the forward fast grant channel is transmitted from more than one BS in the active set, the different BS may operate independently and generate different grant commands. The MS is then required to adjust its Current Grant variable to the minimum of the rate adjustments granted by all the BS in the active set.

[1049] The overhead per MS for continuous request and grant channels is comparable to using the message based channels once every few hundred

ms. Thus, to match the overhead, the rate request and rate grant messages need not be sent more frequently than for example 250 ms per MS, including all services at the MS. Continuous feedback channels may be necessary to manage QoS for services where the maximum delay is less than for example 100 ms. Even for services with bursty arrivals where the queue sizes can vary significantly over tens of ms. If the delay deadlines are greater than 250 ms, then the QoS may be managed through message based request and grant channels. There may be a need for 2-4 bits every 20-40 ms for the continuous reverse rate request channel. These bits indicate to the BS requested modifications to its allocated rate and/or the need for a long grant.

Generally, a framework is disclosed for QoS and resource management on the reverse link, where the packet queues are distributed at the MS and the centralized resource manager is at the BS. In this framework, the responsibility of reverse link QoS management is assigned to the MS while the BS manages the aggregate resource and admission control for negotiated QoS services. The MS provides to the resource manager the required resource (rate or T/P) that it needs to meet its QoS. This is different from previous approaches where the MS provides queue backlog information to the resource manager. The queue backlog information is not sufficient for the resource manager to meet QoS guarantees. Reverse link QoS is managed through a closed loop control of allocated rate or T/P by the MS. The QoS requirements for multiple services (or flows) are merged into a compact representation of resource (rate or T/P). This permits efficient (low communications overhead) for the closed loop control. This permits the MS to determine a required T/P that will allow it to meet QoS requirements of all services. A low complexity mechanism to re-

compute the rate (or T/P) and duration requirements following packet departures, packet arrivals, changes in allocated rate or changes in the link quality. Low-overhead continuous request and grant channels that are consistent with the framework and are suitable to manage QoS for services with maximum delay requirements of less than 100 ms. A process is disclosed for computing T/P requirements for different levels of congestion by using a drop priority for packets. A compact encoding of the rate request channel is disclosed that provides the BS resource manager information to determine level of congestion at the packet queues distributed among the MS. Further, it allows the BS resource manager to make intelligent allocation decisions among the contending MS. Therefore, the BS resource manager operates to allocate the resource among the contending MS, including operation of the closed loop rate control in soft handoff.

Various aspects of the invention may be more apparent by referring to various steps depicted in Fig. 4. Fig. 4 depicts a flow 400 of messages and processing steps at a BS and a MS in the communication system 100. The receiver and transmitter systems 200 and 300 shown in Figs 2 and 3 may operate to perform various steps when incorporated in a respective base station or mobile station in the communication system 100. At step 401, the mobile station determines packets of data for transmission for a number of communication services. At steps 402 and 403, respectively, the mobile station determines a transmission deadline of each of the packets of data and arranges the packets of data in a queue for transmission in accordance with the determined transmission deadline. At steps 405, 406 and 407, respectively, the mobile station determines a data rate for transmission of the packets of data

based on the arrangement of the packets of data in the queue allowing for meeting the transmission deadline for each of the packets of data, determines duration of the determined data rate for transmissions of the packets of data based on the arrangement of the packets of data in said queue, and communicates the data rate and the duration from the mobile station to the base station. At step 408, the base station determines whether available resources allow for allocation at the base station for transmission from the mobile station at the determined data rate and duration. At step 409, the base station communicates acceptance of the determined data rate for transmission of the packets of data from the mobile station. At step 410, the mobile station transmits at the accepted data rate. At step 411, the base station may indicate a congestion level alert to the mobile station that the determined available resources disallow for allocation at the base station for transmission from the mobile station at the determined data rate. At steps 412, 413 and 414, the mobile station drops at least a packet of data of the packets of data in the queue to determine a new queue of packets of data, determines a new data rate for transmission of the new queue of packets of data, the new data rate being lower than the previously determined data rate, and determines a new duration for use of the determined new data rate for transmissions of the packets of data based on the arrangement of the packets of data in the new queue. The flow 400 moves to step 408 to repeat the determination for acceptance or rejection.

Those of skill in the art would further appreciate that the various illustrative logical blocks, modules, circuits, and algorithm steps described in connection with the embodiments disclosed herein may be implemented as

electronic hardware, computer software, or combinations of both. To clearly illustrate this interchangeability of hardware and software, various illustrative components, blocks, modules, circuits, and steps have been described above generally in terms of their functionality. Whether such functionality is implemented as hardware or software depends upon the particular application and design constraints imposed on the overall system. Skilled artisans may implement the described functionality in varying ways for each particular application, but such implementation decisions should not be interpreted as causing a departure from the scope of the present invention.

[1050] The various illustrative logical blocks, modules, and circuits described in connection with the embodiments disclosed herein may be implemented or performed with a general purpose processor, a digital signal processor (DSP), an application specific integrated circuit (ASIC), a field programmable gate array (FPGA) or other programmable logic device, discrete gate or transistor logic, discrete hardware components, or any combination thereof designed to perform the functions described herein. A general-purpose processor may be a microprocessor, but in the alternative, the processor may be any conventional processor, controller, microcontroller, or state machine. A processor may also be implemented as a combination of computing devices, e.g., a combination of a DSP and a microprocessor, a plurality of microprocessors, one or more microprocessors in conjunction with a DSP core, or any other such configuration.

[1051] The steps of a method or algorithm described in connection with the embodiments disclosed herein may be embodied directly in hardware, in a software module executed by a processor, or in a combination. A software

module may reside in RAM memory, flash memory, ROM memory, EPROM memory, EEPROM memory, registers, hard disk, a removable disk, a CD-ROM, or any other form of storage medium known in the art. An exemplary storage medium is coupled to the processor such that the processor can read information from, and write information to, the storage medium. In the alternative, the storage medium may be integral to the processor. The processor and the storage medium may reside in an ASIC. The ASIC may reside in a user terminal. In the alternative, the processor and the storage medium may reside as discrete components in a user terminal.

[1052] The previous description of the preferred embodiments is provided to enable any person skilled in the art to make or use the present invention. The various modifications to these embodiments will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other embodiments without the use of the inventive faculty. Thus, the present invention is not intended to be limited to the embodiments shown herein but is to be accorded the widest scope consistent with the principles and novel features disclosed herein.